iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 6 **Issue:** I **Month of publication:** January 2018

DOI: http://doi.org/10.22214/ijraset.2018.1175

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor :6.887
Volume 6 Issue I, January 2018- Available at www.ijraset.com

Analysis of Video Summarization Techniques

Parminder Kaur[1], Dr. Rakesh Kumar [2]
[1]*Research Scholar, Department of Computer Science and Applications, Kurukshetra University, Kurkshetra*
[2]*Professor, Department of Computer Science and Applications, Kurukshetra University, Kurukshetra*

Abstract: The increased amount of video data makes the browsing, retrieval and processing of video data a difficult task. Video Summarization is proposed as a solution to this problem. There are two types of video summaries static and dynamic. This paper categorizes various types of video summarization techniques studied, on deferent basis.
Keywords: clustering, features, key frames, skimming, summarization.

I.INTRODUCTION

In the recent years there has been a tremendous increase in the technology. With the increase in the affordability and availability of low price and high quality video recording devices the amount of video data has grown by leaps and bounds. As per a survey report since 2014 Youtube usage has almost tripled and approximately 400 hours of new video is uploaded every minute. This figure is only from the most popular video search engine, if facts and figures from other sources such as real time video feeds from surveillance cameras are collected then these facts would be more astounding. The processing of such huge amount of data itself is a challenge.

A possible solution to this problem is video summarization; it is often called as video abstraction. Video summarization is a concise and meaningful representation of a video. It not only reduces the amount of processing data but also makes the processing simpler task. The video is summarized either by eliminating the redundant video data or by selecting the salient contents. There are two types of video summaries- static summaries and dynamic summaries as shown in Figure 1. The static summaries are constructed by selecting the salient frames or the representative frames of the video it is also known as key frame extraction. Dynamic summaries also called video skims; it is a segment of the video itself and represents the important contents. Video skims possess higher level meaning and they are similar to trailers of movies. Skims are visually more appealing and often offer greater understanding of the situation to the viewer. Dynamic video summarization is a complex task and requires different modules for handling deferent type of information.

A video is a multimedia sequence of images which may have audio also. For the purpose of processing, a long duration video may be decomposed into small segments on the basis scenes or shots. The collection of semantically and temporally related groups of elements of a video that convey a higher level meaning is called a scene. There are many techniques for video summarization which are based on scene level decomposition. A shot is defined as a sequence of actions captured by a single camera with no major changes in the visual content. It represents a physical concept and usually shot boundary based techniques are used for video summarization. The shot boundary is identified abrupt and gradual (fade in, fade out) changes in the frames. And the representative frame from each shot is selected to construct the summary of the entire video.

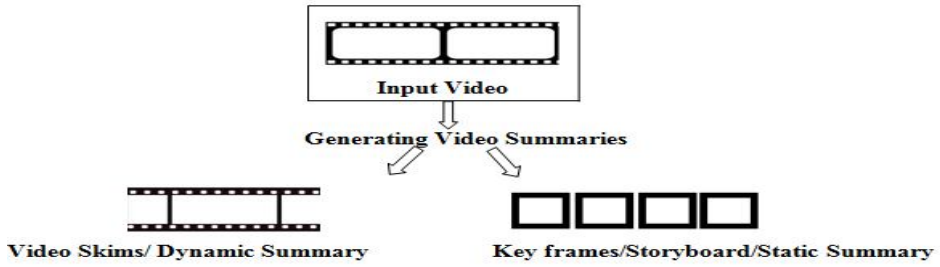


Figure1 Types of Video Summarization

This paper is organized as follows. In Section II, various summarization techniques available in literature are studied and categorized on different basis, evaluation techniques are analyzed in Section III. And finally conclusions are provided in Section IV.

II. VIDEO SUMMARIZATION TECHNIQUES

There are many techniques for video summarization with different basis of classification. The existing techniques can however be classified on the basis of features used as- Low Level Feature based Video Summarization and High Level Feature based Video Summarization, Single Feature based Video Summarization and Multiple Feature based Video Summarization; on the basis of technique used - Clustering based Video Summarization and Non-Clustering based Video Summarization and the recent techniques. The classification of the techniques is described below.

A. Feature based Video Summarization

Based on the type and the number of features used for determining the representation of a frame and identifying the visually important contents the video summarization techniques can be categorized as below.

1) Low Level Feature based Video Summarization: The techniques which were used earlier for video summarization were based on low level features such as color, texture, motion for extracting the visually important and relevant information from the video. The most commonly used low level feature is color. It is easy compute. Color histograms are obtained for each frame and based on histogram difference's or histogram intersection, frames can be compared. Image/Color histogram refers to the probability mass function of the intensities and it is computed by counting the number of pixels of same color. Although RGB [14] [4] color model is the most commonly used color model but HSV [14] [2] color model has closeness to the human perception, there-fore it is more preferred than RGB color model. RGB color model is based on neuro-physiology [1] where as HSV is an example of a psychologically [1] inspired color model. Color histograms are insensitive to small camera motions and are easy to compute. While comparing two frames by using color histograms, one may encounter a situation in which two frames are entirely different but there histograms are same because of the similar color distribution of the frames. Texture is also a commonly used low level feature. There are many techniques for texture extraction, mostly wavelet trans-form is used for texture analysis but in [14] Discrete Haar Wavelet Transforms is used whereas Daubechies wavelet transform s used in [11].

Motion is considered as one of the important features for capturing the visually interesting elements. The most common methods for computing motion are motion histograms, optical flow analysis. Motion histograms [3] are analogous to the color histograms. Optical Flow Analysis is the most simple and common technique for computing the motion. Optical Flow analysis is based on the brightness consistency constraint, which states that the brightness remains the same which limits the applicability of the method to an environment which is not effected by illumination changes.

The benefit of using low level features is that they are easy to compute which makes these features a first choice for the applications which require quick response and the scenes are not complex. But there are many complex scenes having important information which remain unnoticed if low level features are used. So with the advancing technologies the quality of the videos and the complexity of scenes also increased and thereby moved the focus from low level features to high level features for identifying visually important contents from a video.

2) High Level Feature based Video Summarization: Low level features are easy to compute but they are not close to the semantics however high level features such as object recognition, event detection, face detection gesture detection, emotion detection etc aids in better understanding of the contents of the video. Lot of work is done in the field of event detection in sports videos to create summary or abstracts of some special events of the game like goal in football, bout in wrestling etc. The model proposed in such cases is trained to recognize such events. A technique for generating video summaries for the user generated videos is proposed in [17]. It makes use of the emotions and determines the semantically important contents and by making the use of these features summary is generated. In [12] static summaries are generated for videos captured by wearable cameras. An object driven approach is used to find out the important contents it makes use of region features, object features for compute the relevance of the frame contents.

High level feature based video summarization approaches are application specific and very useful for the applications that are very sensitive to the contents. Despite its ad-vantages high level feature based summarization techniques are time consuming and they are computationally expensive.

3) Single Feature based Video Summarization: The simplest techniques for video summarization makes use of single feature usually either color or motion as the only descriptor. Using a single feature simplifies the computation task and saves time also. Most of the earlier work done was based on color models only. There are many studies that make use of motion as a sole feature to find out the relevant contents in a video. In an analogy to color histogram motion histograms [3] were also proposed. A body in motion first accelerates then decelerates after a point, using this law and with an assumption that the particular point where the

motion starts decelerating is the point of interest and the triangle between the acceleration and deceleration represents the perceived motion energy, with this presumption PME(Perceived Motion Energy) model was developed [13]. Using PME Model motion patterns are modeled and after that the key frames are extracted.

Single features based video summarization techniques are computationally simple and easy but they may not always provide reliable results.

4) Multiple Features based Video Summarization: With the increasing complexity of the video contents it was realized that a single descriptor was not sufficient for evaluating the importance and relevance of the contents. Techniques which make use of multiple features were proposed. [5] is based of the three features representation for a frame-color histograms, edge direction histograms and texture statistics. [4] aggregated the correlation of RGB color channels, moments of inertia and color histograms using an adaptive formula for generating the video summaries. While combining multiple features some of the techniques made use of a formula which is a linear function; in other words it was assumed that all the features are equally important. But there are many real life situations where not all the features are equally important and may fail to capture the important contents. So techniques in which each feature is assigned some weight based on its relative importance were developed. [9] used CMM(Convex Mixture Models) for computing the weights of various visual descriptors whereas [4] assigned predefined weights to different features, these weights were empirically determined.

User Attention Models[10] [11] is a computational model which imitates the human visual system is based on the fact that the human eye is sensitive to both color contrast and motion, so these models combine these features to compute the attention curves.

Exploiting multiple features for evaluating the saliency of the contents makes the summarization approach more effective. Using multiple features explores and exploits the contents from each and every aspect but at the same time the feasibility of the technique should be evaluated because in certain cases it is not cost efficient to use multiple features for evaluation.

B. Technique based Video Summarization

The traditional approaches to video summarization are based on clustering techniques but there are some non clustering based techniques which can also be used to summarize the video. Both categories are described below.

1) Shot Detection based Video Summarization: A video is a continuous sequence of images which is sometimes segmented into small shots based on the identification of the shot boundaries. A shot boundary is basically identified either by abrupt changes in the contents of the video or by the gradual (fades in, fade out) changes in the contents. Once shots are detected based on some features, a frame from the shot is selected as the representative of the shot. All the representative frames from the shots contribute to the summary of the video. There are some techniques which make use of frame differences (color histograms) for detecting the shot boundaries and at later stage representative frames were selected [5]. Frame differences are usually used to detect the shot boundaries. Frame differences can be computed by using any feature but the most commonly used feature which is easily computable is color histograms, [7] used SIFT point distribution histograms for computing the frame differences and generating the video summaries.

Shot boundary detection algorithms are simple and easy to implement but they have restrictive applicability. These algorithms are dependent on threshold values. If a very small value is set it would lead a large number of shots and thereby resulting in a long summary. Best results of shot based video summarization can be obtained on professionally generated videos but they can't be used for user generated videos or egocentric videos. User generated videos and egocentric videos are continuous in nature and lacks abrupt changes in the contents.

2) Clustering based Video Summarization: The basic idea of clustering based approaches is to consider each frame as a point and by applying the clustering algorithm the cluster representatives are selected as key frame. There are many clustering algorithms - k means [2], k mediods, guarded zone clustering, spectral clustering [8], MST(Minimum Spanning Tree) clustering [15] etc which were used in the literature for the purpose of video summarization. In clustering based key frame extraction techniques (static video summarization) usually the key frames generated are not in the chronological order, which often creates a semantic gap for understanding the sequence of events. So as a solution to this problem there is some work done to preserve the temporal sequence of the video, one such time constraint cluster algorithm was developed in [11].

3) Non- Clustering based Video Summarization: There are many thresholding based techniques for selecting the exemplar frames of a video. A thresholding based technique for sequential selection of key frames is described in [16]. Sum of absolute differences of histograms of frames is computed and the frames for which the sum of absolute difference is greater than the predefined threshold are selected as key frames. The efficiency of the threshold based techniques largely determined by threshold value. If the threshold value is very low then the number of key frames will be large and if it is very high there may be cases that not even a single frame

qualifies as a key frame. Moreover in order to empirically determine the threshold value, experiments with all possible type of data sets need to be per-formed. Curvature frame difference based techniques was used in [5]. There are techniques which were based on pixel wise (pixel based frame differences) comparisons of the frames. Although it is easy to compute the pixel based frame differences but the approach doesn't provide reliable results when the camera angle changes. Later it was found that a possible solution to this problem can be dividing the frame into blocks and instead of using pixel based frame difference blocks based frame differences should be computed. [10] used Visual Attention Model for finding out the visually salient contents. In the proposed User Attention Model, static dynamic and se-mantic attentions curves are computed and for summarizing the video all the curves are combined to obtain the global attention curve. In this approach the crests of the global attention curve are considered as key frames that represent the visually salient contents of the video. One of the biggest advantages of a User Attention Model is that it doesn't require the complete understanding of the visual con-tents. It captures the cognitive functions responsible for identifying visually different information. Using the Visual Attention Model even doesn't require the thorough investigation/study of the Visual Attention System. Some of the proven facts and conclusions can be used to model the visual attention. Although there are many advantages of Visual Attention Model based video summarization techniques but there are some disadvantages of these models. They are complex as they are analogous to human visual system moreover there are various attention models which are later combined to calculate the overall visually attractive contents. These attention models create the complexity. This complexity can however be reduced by using parallel processing.

C. Recent Techniques

SIFT(Scale Invariant Feature Transform) algorithm is not sensitive to change in illumination, scale and noise. [7] used SIFT for extracting the key points of the frames and then represented each frame by the SIFT point histograms, shot boundaries were detected and then key frames were extracted from each frame. The advantage of the SIFT algorithm is that it is not susceptible to noise, change in illumination, rotation and scale (zooming in or out). But de-spite having this advantage one of the major disadvantage is its computational complexity moreover it needs training of the system. These disadvantages make this algorithm a less likely choice for real time applications. The video summarization techniques studied were mostly based on the notion of using only the imagery features but there are other non visual sources of information such as audio, geo location data, textual data from social networking sites etc which when combined with imagery data can be very helpful for extracting the visually important contents of the video. One such framework which learns from multiple sources for the task of video summarization was proposed in [18]. Though learning from multiple sources sounds interesting but there are many hurdles in this technique. The on time availability of the non imagery data and the effect of noise on the audio data captured from public places furthers increases the computational complexity. An event based sports video segmentation technique that utilizes the webcast text data along with the timestamps of the events is used for segmenting the events was developed in [6].

III. EVALUATION TECHNIQUES

The video summaries generated needs to be evaluated for proving their superiority over the other techniques. However, this area of video analysis doesn't have standard performance indicators and evaluation techniques. Lack of standard techniques makes the process more challenging. Evaluating the video summaries is a highly subjective task but there are some methods for evaluating the summaries. The existing techniques can be categorized as:

A. Subjective Evaluation

The subjective form of evaluating the summaries has a very high dependency on humans. The most basic approach was the visual assessment of the summaries but due to the lack of practicality of the approach some other techniques were developed. User summaries were generated by the users by observing the videos. The group of users who were assigned the task of watching the videos and then summarizing the videos are selected from different backgrounds and age groups to avoid biasness. The user group usually annotates the videos and the annotations were compared with the summaries in order to evaluate them. Apart from user generated summaries and annotations there are some other parameters for evaluation like informativeness and pleasantness of the contents of the automatically generated summary.

B. Objective Evaluation

Without any objective results of the evaluation procedure, it becomes difficult to assess the automatically generated summary. There are some techniques which try to use objective evaluation methods; a very common and widely used technique is CUS (Comparison of User Summaries) [4]. CUS involves human intervention but at the same time it provides numerical results also. In CUS a group of users generate the summaries; the summaries generated by the user are compared with the automatically generated summaries. The frames from both user generated summary and automatically generated are compared by computing the Manhattan distance between them or by using the difference of the color histograms. Other techniques which were used in shot boundary detection based summarization approaches are- shot reconstruction degree, fidelity and compression ratio [16]. Shot reconstruction degree identifies the extent to which a shot can be constructed by using the key frames and a frame interpolation algorithm. Compression ratio is the ratio of the no of key frames selected to the number of frames in the shot; it studies the compactness of the shot. Fidelity [16] measures the distance between the key frame and the other frames of the shot it uses Housdorff distance.

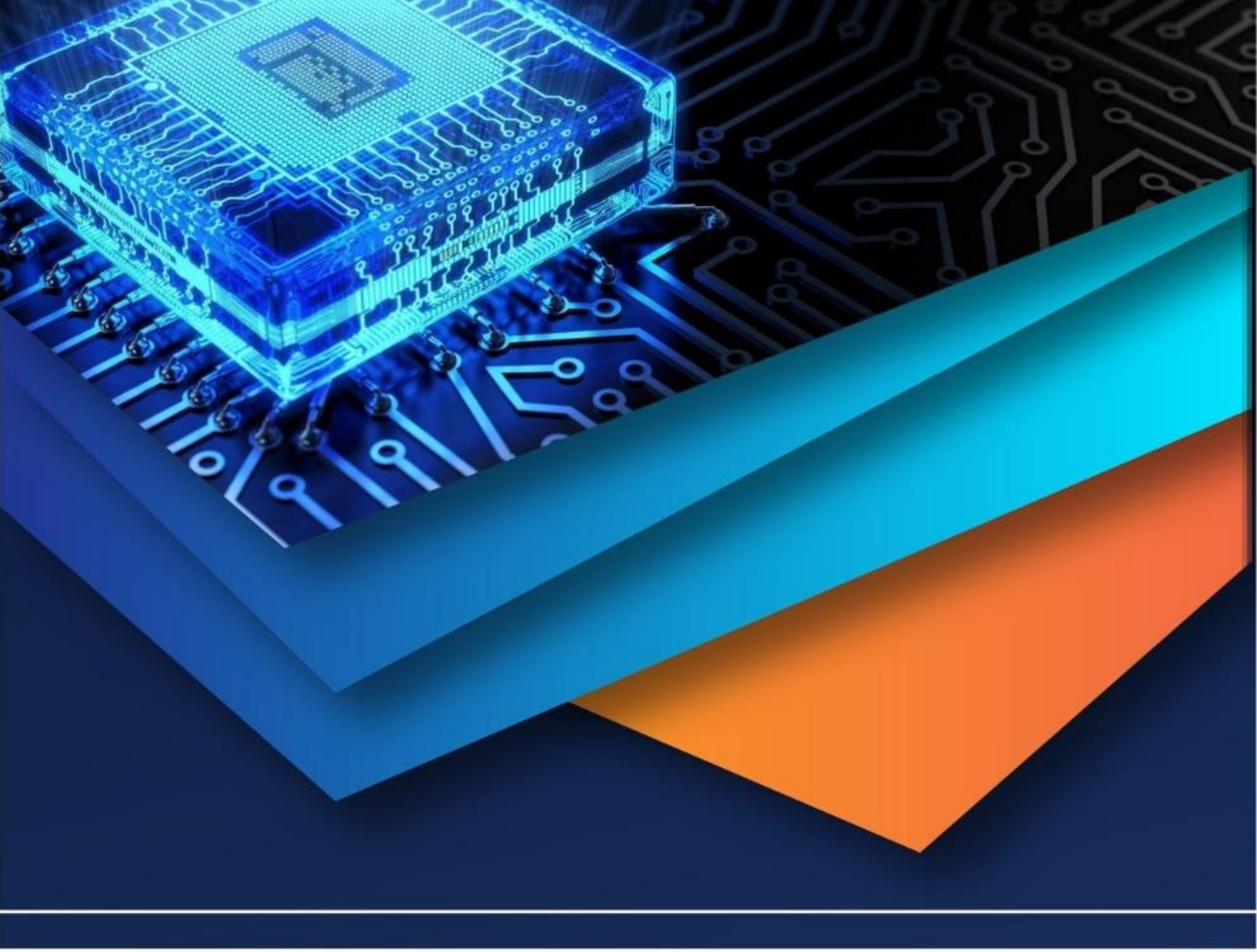
IV. CONCLUSION

The recent advancements in the field of video analytics have driven the need for automatic video summarization. There are many techniques for video summarization, which were studied and categorized. It is observed that not all the summarization techniques fit well in each and every situation. Some of the techniques (Low level feature based) are good for real time applications as they are computationally simple and fast; where as some techniques(High level feature based, User attention model based) are particularly suitable for applications that require precise and accurate results regardless of the time taken for producing the summary(e.g. surveillance applications). Each technique has its own merits and demerits but the need for a technique which is independent of the application is realized. Secondly there is lack of standard evaluation techniques, earlier user generated summaries were used to evaluate the automatic generated summary later shot reconstruction degree, fidelity were proposed and used.

REFERENCES

[1] Francesco Camastra and Alessandro Vinciarelli. “Machine learning for audio, image and video analysis". In: Advanced Information and Knowledge Processing (2008), pp. 83-89.
[2] Sandra Eliza Fontes De Avila et al. “VSUMM: A mechanism designed to produce static video summaries and a novel evaluation method". In: Pattern Recognition Letters 32.1 (2011), pp. 56-68.
[3] Yining Deng and BS Manjunath. “Content based search of video using color, texture, and motion". In: Image Processing, 1997. Proceedings, International Conference on. Vol. 2. IEEE. 1997, pp. 534-537.
[4] Naveed Ejaz, Tayyab Bin Tariq, and Sung Wook Baik. “Adaptive key frame extraction for video summarization using an aggregation mechanism". In: Journal of Visual Communication and Image Representation 23.7 (2012), pp. 1031-1040.
[5] Ciocca Gianluigi and Schettini Rai-mondo. “An innovative algorithm for key frame extraction in video summarization". In: Journal of Real-Time Image Processing 1.1 (2006), pp. 69-88
[6] Gyeong-June Hahm and Keeseong Cho. “Event-based sport video segmentation using multimodal analysis". In: International Conference on Information and Communication Technology Convergence (ICTC), 2016. IEEE. 2016, pp. 1119-1121
[7] Rachida Hannane et al. “An efficient method for video shot boundary detection and keyframe extraction using SIFT-point distribution histogram". In: International Journal of Multimedia Information Retrieval 5.2 (2016),89-104
[8] Antonis I Ioannidis, Vasileios T Chasanis, and Aristidis C Likas. “Key-Frame Extraction Using Weighted Multi-view Convex Mixture Models and Spectral Clustering". In: 22nd International Conference on Pattern Recognition (ICPR), 2014. IEEE. 2014, pp. 3463-3468
[9] Antonis Ioannidis, Vasileios Chasanis, and Aristidis Likas. “Weighted multiview key-frame extraction". In: Pattern Recognition Letters 72 (2016), pp. 52-61
[10] Hugo Jacob et al. “A video summarization approach based on the emulation of bottom-up mechanisms of visual attention". In: Journal of Intelligent Information Systems (2017), 1-19.
[11] Jie-Ling Lai and Yang Yi. “Key frame extraction based on visual attention model". In: Journal of Visual Communication and Image Representation 23.1 (2012), pp. 114-125
[12] Yong Jae Lee, Joydeep Ghosh, and Kristen Grauman. “Discovering important people and objects for ego-centric video summarization". In: Conference on Computer Vision and Pattern Recognition (CVPR), 2012. IEEE. 2012, pp. 1346-1353.
[13] Tianming Liu, Hong-Jiang Zhang, and Feihu Qi. “A novel video key-frame extraction algorithm based on perceived motion energy model". In: IEEE transactions on circuits and systems for video technology 13.10 (2003), pp. 1006-1013.
[14] Karim M Mahmoud, Mohamed A Ismail, and Nagia M Ghanem. “Vscan: an enhanced video summarization using density based spatial clustering". In: International Conference on Image Analysis and Processing. Springer. 2013, pp. 733-742.
[15] Rameswar Panda, Sanjay K Kuanar, and Ananda S Chowdhury. “Scalable video summarization using skeleton graph and random walk". In: 22nd International Conference on Pattern Recognition (ICPR), 2014. IEEE. 2014, pp. 3481-3486.
[16] CV Sheena and NK Narayanan. “Key-frame Extraction by Analysis of Histograms of Video Frames Using Statistical Methods". In: Procedia Computer Science 70 (2015), pp. 36- 40.
[17] Baohan Xu, Xi Wang, and Yu Gang Jiang. “Fast Summarization of User Generated Videos: Exploiting Semantic, Emotional, and Quality Clues". In: IEEE MultiMedia 23.3 (2016), pp. 23-33.

[18] Xiatian Zhu, Chen Change Loy, and Shaogang Gong. “Learning from multiple sources for video summarisation". In: International Journal of Computer Vision 117.3 (2016), pp. 247-268.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY